Section 9.1 (Part 1 of 2)
Understanding Test of Hypothesis Basics

KEY CONCEPTS

INFERENCE TESTS:

□ Confidence Interval (Chapter 8): ESTIMATES ALL OF THE PLAUSIBLE VALUES FOR A POPULATION PARAMETER ← μ, P

□ vs. Significance Tests (Chapters 9-12):
ASSESS THE EVIDENCE PROVIDED ABOUT SOME CLAIM CONCERNING A Population

DEFINE HYPOTHESES: → ALWAY REFER TO THE POPULATION

□ State the parameter of interest YOU MUST CLEARLY LABEL < P = / μ =

□ Null Hypothesis (H_0) NO DIFFERENCE. THE CLAIM WE ARE SEEKING EVIDENCE AGAINST
$\bar{x}$ NOT

□ Alternative Hypothesis (H_A) THE CLAIM WE HOPE OR SUSPECT TO BE TRUE INSTEAD OF NULL HYPOTHESIS (H_0)

STATISTICAL INFERENCE:

□ ~~Statistical Test~~ → CLAIM IS A STATEMENT ABOUT A PARAMETER (p or μ)

□ 2 Outcomes of a Statistical Test ① FAIL TO REJECT the NULL HYPOTHESIS
② REJECT NULL HYPOTHESIS IN FAVOR OF THE ALTERNATE

□ P-Value
the probability assuming Ho is true – small p-value reject Ho
Small pvalue suggests the observed result is unlikely to occur if Ho is true.

□ Statistically Significant

1) SIGNIFICANCE LEVEL (α=.05)

2) STATISTICALLY SIGNIFICANT
* when the p-value issmaller than the preselected α.
Reject Ho.

NEVER! ACCEPT Ho
"we will never know the true pop parameter.

Introduction

- **There are 2 types of statistical inference** - Confidence intervals and *Significance Tests*.
  1. **Confidence intervals** estimate a population parameter (μ or p)
  2. **Significance Tests** assess the evidence provided by data about some claim concerning a population.
     - **Significance Tests**
       - formal procedure for comparing observed data with a claim (also called a hypothesis) whose truth we want to assess.
       - The claim is a statement about a parameter, like the population proportion *p* or the population mean μ.
       - We express the results of a significance test in terms of a probability that measures how well the data and the claim agree.

In this chapter:

- **we'll learn the underlying logic of statistical tests,**

  how to perform tests about population proportions and population means, and
- **how tests are connected to confidence intervals.**

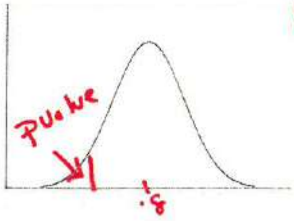
■ Stating Hypotheses

See Poster "TESTING Hypotheses"

① STATE PARAMETER

$p =$

$\mu =$



EX #1

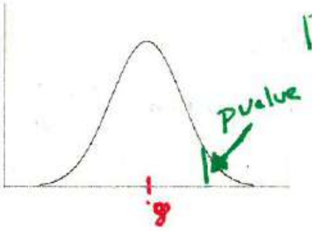
$H_0: p = .80$ (shoots 80%)

$H_a: p < .80$ (shoots less than 80%)

② $H_0: p =$

$H_0: \mu =$

NO DIFFERENCE



EX #2

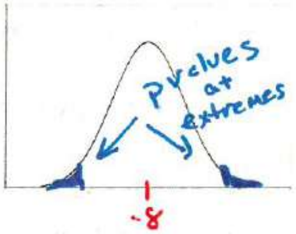
$H_0: p = .80$ (shoots 80%)

$H_A: p > .80$ (shoots more than 80%)

③ $H_A: p$

μ

claim we hope or suspect to be true instead of H_0



EX #3 2-SIDED TEST

$H_0: p = .80$ (shoots 80%)

$H_A: p \neq .80$ (does not shoot 80%)

A significance test starts with a careful statement of the claims we want to compare.

- ✓ The first claim is called the **null hypothesis**. Usually, the null hypothesis is a statement of "no difference."
- ✓ The claim we hope or suspect to be true instead of the null hypothesis is called the **alternative hypothesis**

In any significance test, the null hypothesis has the form

$$H_0 : \textbf{parameter = value}$$

The alternative hypothesis has one of the forms

$$H_a : \textbf{parameter < value}$$
$$H_a : \textbf{parameter > value}$$
$$H_a : \textbf{parameter} \neq \textbf{value}$$

To determine the correct form of H_a, read the problem carefully.

✓ **Hypotheses always refer to a *population***, not to a sample. Be sure to state H_0 and H_a in terms of *population parameters*.

✓ It is ***never*** correct to write a hypothesis about a sample statistic, such as $\hat{p}$ or $\bar{x}$

Definition:

The alternative hypothesis is **one-sided** if it states that a parameter is *larger than* the null hypothesis value or if it states that the parameter is *smaller than* the null value.

It is **two-sided** if it states that the parameter is *different* from the null hypothesis value (it could be either larger or smaller).

Example #2: The Basketball Player

The Reasoning of Significance Tests

> **EXAMPLE:** Suppose a basketball player claimed to be an 80% free-throw shooter. To test this claim, we have him attempt 50 free-throws. He makes 32 of them. His sample proportion of made shots is 32/50 = 0.64. What can we conclude about the claim based on this sample data?

1. **What is the population parameter we want to test?**

   p = the longrun proportion of free throws

2. **What hypothesis do we want to test (in symbols and words)?**

   H_0: p = .80 (players true foul shooting is 80%)

   H_A: p < .80 (player shoots less than 80%)

3. **What evidence do we have (assume conditions of random independent and normal are met)?**

   ① CONFIDENCE INTERVAL:

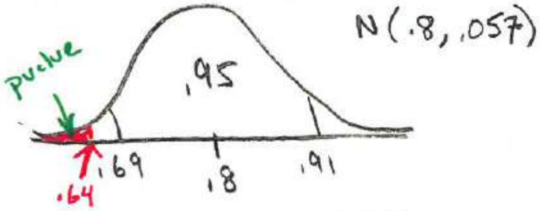
   WE KNOW HOW TO FIND 95% CI for p

   $\mu_{\hat{p}} = .80$

   $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{.8(.2)}{50}} \approx .057$

   CI: $.8 \pm 1.96(.057)$

   $.8 \pm .11$ $(.69, .91)$



② SIGNIFICANCE TEST

pvalue = $P(p \leq .64) =$

$P(z \leq \frac{.64-.80}{.057} = -2.81)$

$P(z \leq -2.81) = .0021$

normalcdf (-E99, -2.81, 0, 1

4. **Do we have enough evidence to reject our null hypothesis?**

   ① CI Gives a range of plausible value for p, Since .64 falls outside the range we have evidence to reject Ho and believe he is NOT an 80% FT shooter.

Read highligted conclusion

In the free-throw shooter example,

1. Population parameter: *p* is the long-run proportion of made free throws.
2. Our hypotheses are:

   H0 : *p* = 0.80 (the basketball player does have a 80% foul shooting %) NO DIFFERENCE CLAIM

   Ha : *p* < 0.80 (the basketball player shoots less than 80% foul shooting)

   CI for $\hat{p} = .64$

   CI: $.64 \pm 1.96(.068)$

   $.64 \pm .13$

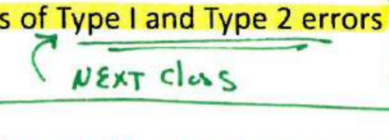
   $(.51, .77)$

3. Draw a normal curve with mean phat=.8 and critical value .64

   phat= 0.80 (our point estimate of the population $\hat{p} = .64$

   SE= sqrt(pq/n) =sqrt(.8*.2/50)=.0567 $\sigma_{\hat{p}} = \sqrt{\frac{(.64)(.36)}{50}} = .068$

   N(.8,.0567) → normalcdf(-e99, .64, 8, .0567) =.0023



.80 IS NOT IN CI provides evidence he is not an 80% FT shooter.

4. We found a p-value of .002~~3~~4 which is very small and it is unlikely this would occur by chance

   So we would reject the null hypthesis in favor of the alternate hypothes that he shoots less than 80%

   But we could have made a mistake and that leads to discussions of Type I and Type 2 errors

   NEXT class

- *Statistical tests deal with claims about a population.*
- *Tests ask if sample data give good evidence against a claim. A test might say, "If we took many random ples and the claim were true, we would rarely get a result like this."* **SEE GRAPH**
- *You can say how strong the evidence against the player's claim is by giving the probability that he would make as few as 32 out of 50 free throws if he really makes 80% in the long run.*
- **The observed statistic is so unlikely if the actual parameter value is *p* = 0.80 that it gives convincing evidence that the player's claim is not true.**

The Reasoning of Significance Tests

Based on the evidence we might conclude the player's claim is incorrect.

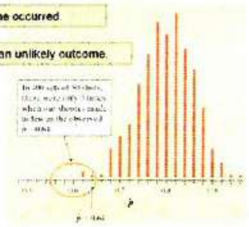
In reality, there are two possible explanations for the fact that he made only 64% of his free throws.

1) The player's claim is correct ($p = 0.8$), and by bad luck, a very unlikely outcome occurred.

2) The population proportion is actually less than 0.8, so the sample result is not an unlikely outcome.

Basic Idea

An outcome that would rarely happen if a claim were true is good evidence that the claim is not true.



■ Example #3: Studying Job Satisfaction

■ ***Stating Hypotheses***

EXAMPLE: Does the job satisfaction of assembly-line workers differ when their work is machine-paced rather than self-paced? One study chose 18 subjects at random from a company with over 200 workers who assembled electronic devices. Half of the workers were assigned at random to each of two groups. Both groups did similar assembly work, but one group was allowed to pace themselves while the other group used an assembly line that moved at a fixed pace. After two weeks, all the workers took a test of job satisfaction. Then they switched work setups and took the test again after two more weeks. The response variable is the difference in satisfaction scores, self-paced minus machine-paced.

a) Describe the parameter of interest in this setting.

μ = difference in satisfaction score (SELF DACED — MACHINE PACED)

b) State appropriate hypotheses for performing a significance test. (in symbols and words)

H_0: $\mu = 0$ No difference in job satisfaction scores

H_A: $\mu \neq 0$ There is a difference in job satisfaction scores

"workers could either be more satisfied or less satisfied"

a) The parameter of interest is the mean μ of the differences (*self-paced minus machine-paced*) in job satisfaction scores in the population of all assembly-line workers at this company.

b) Because the initial question asked whether job satisfaction differs, the alternative hypothesis is two-sided; that is, either $\mu < 0$ or $\mu > 0$. For simplicity, we write this as $\mu \neq 0$. That is,

***H_0*: $\mu = 0$** NO differences (*self-paced minus machine-paced*) in job satisfaction scores
***H_a*: $\mu \neq 0$** There is a difference in job satisfaction scores

- **Example #3:** Studying Job Satisfaction (continued)
- ***Interpreting Null Hypothesis and P-Value***

For the job satisfaction study, the hypotheses are

$$H_0: \mu = 0$$
$$H_a: \mu \neq 0$$

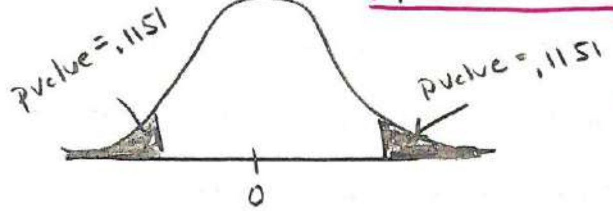
Data from the 18 workers gave $\bar{x} = 17$ and $s_x = 60$. That is, these workers rated the self - paced environment, on average, 17 points higher. Researchers performed a significance test using the sample data and found a P - value of 0.2302.

c) What null hypothesis means in this setting when it is true

IF H_0 IS TRUE, THEN THE WORKERS DO NOT FAVOR ONE WORK ENVIRONMENT OVER THE OTHER, ON AVERAGE.

d) Interpret the *P*-value in context.

pvalue = .2302 → large pvalue → FAIL TO REJECT H_0 in favor of H_a.



Conclusion: We fail to reject H_0, with an average difference of 17 or more points between the 2 work environments would happen 23% OF THE TIME JUST BY CHANCE IN RANDOM SAMPLES OF 18 workers. THIS is NOT CONVINCING EVIDENCE AGAINST H_0.

Small pvalue → Reject H_0 → Conclude H_a (in context)

large pvalue → FAIL TO REJECT H_0 → Can not conclude H_A (in context)

c) In this setting, H_0: μ = 0 says that the mean difference in satisfaction scores (*self-paced - machine-paced*) for the entire population of assembly-line workers at the company is 0.

If H_0 is true, then the workers don't favor one work environment over the other, on average.

d) P-Value in context

Because the alternative hypothesis is two - sided, the P - value is the probability of getting a value of $\bar{x}$ as far from 0 in either direction as the observed $\bar{x} = 17$ when H_0 is true. That is, an average difference of 17 or more points between the two work environments would happen 23% of the time just by chance in random samples of 18 assembly - line workers when the true population mean is $\mu = 0$.

An outcome that would occur so often just by chance (almost 1 in every 4 random samples of 18 workers) when H_0 is true is not convincing evidence against H_0.
We fail to reject H_0: μ = 0.

e) Definition of P-Value:

- ✓ **The p-value measures the strength of the evidence against a null hypothesis.** The *P*-value is the probability of observing a sample result as extreme or more extreme in the direction specified by H_a just by chance when H_0 is actually true.
- ✓ **Small p-values are evidence against Ho. So we would reject the null hypothesis.**
- ✓ **Large p-values fail to give convincing evidence against Ho because they say that the observed result is likely to occur by chance when H0 is true. So we would fail to reject the null hypothesis.**
- ✓ **POINT TO BE MADE: (draw a diagram) Because the alternative hypothesis is a 2-sided test, the *P*-value is the probability of observing a sample result at either extreme from the null hypothesis Ho=0.**

■ Example #4: Better Batteries

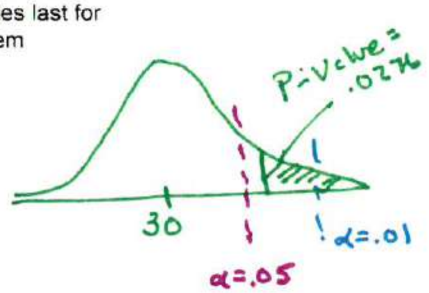
■ *Statistically Significance at level α*

A company has developed a new deluxe AAA battery that is supposed to last longer than its regular AAA battery. However, these new batteries are more expensive to produce, so the company would like to be convinced that they really do last longer. Based on years of experience, the company knows that its regular AAA batteries last for 30 hours of continuous use, on average. The company selects an SRS of 15 new batteries and uses them continuously until they are completely drained. A significance test is performed using the hypotheses

IDENTIFY TOH!
IN AK

$$H_0 : \mu = 30 \text{ hours}$$
$$H_a : \mu > 30 \text{ hours}$$

where μ is the true mean lifetime of the new deluxe AAA batteries. The resulting *P*-value is 0.0276.



a) What conclusion can you make for the significance level $\alpha = 0.05$?

pvalue =.0276 < α=.05 → REJECT H_0

The sample result is statistically significant at the 5% level. We have sufficient evidence to Reject Ho, and conclude that the Company's deluxe AAA batteries last longer than 30 hours, on average.

b) What conclusion can you make for the significance level $\alpha = 0.01$?

p value =.0276 ≥ α=.01 → FAIL TO REJECT Ho

The sample result is NOT statistically significant at the 1% level.

WE do NOT HAVE ENOUGH EVIDENCE AND FAIL TO REJECT Ho. THEREFORE WE CAN NOT CONCLUDE THAT THE deluxe AAA batteries last longer than 30 hours, ON AVERAGE

The final step in performing a significance test is to draw a conclusion about the competing claims you were testing. We will make one of two decisions based on the strength of the evidence against the null hypothesis (and in favor of the alternative hypothesis) -- **reject H_0 or fail to reject H_0.**

When we use a fixed level of significance to draw a conclusion in a significance test,

***P*-value < α → reject H_0 → conclude H_a (in context)**

***P*-value ≥ α → fail to reject H_0 → cannot conclude H_a (in context)**

If α = 0.05 :

Since the *P*-value, 0.0276, is less than $\alpha = 0.05$, the sample result is statistically significant at the 5% level. We have sufficient evidence to **reject *H0*** and conclude that the company's deluxe AAA batteries last longer than 30 hours, on average.

If α = 0.01 :

Since the *P*-value, 0.0276, is greater than $\alpha = 0.01$, the sample result is not statistically significant at the 1% level. **We do not have enough evidence** (fail) **to reject H_0 in this case. therefore, we cannot conclude that the deluxe AAA batteries last longer than 30 hours, on average.**

Significance Tests: The Basics

Summary

- ✓ A **significance test** assesses the evidence provided by data against a **null hypothesis H_0** in favor of an **alternative hypothesis H_a**.

- ✓ The hypotheses are stated in terms of population parameters. Often, H_0 is a statement of no change or no difference. **H_a** says that a parameter differs from its null hypothesis value in a specific direction (**one-sided alternative**) or in either direction (**two-sided alternative**).

- ✓ The reasoning of a significance test is as follows. Suppose that the null hypothesis is true. If we repeated our data production many times, would we often get data as inconsistent with H_0 as the data we actually have? If the data are unlikely when H_0 is true, they provide evidence against H_0 .

- ✓ The ***P*-value** of a test is the probability, computed supposing H_0 to be true, that the statistic will take a value at least as extreme as that actually observed in the direction specified by H_a .

- ✓ Small *P*-values indicate strong evidence against H_0 . To calculate a *P*-value, we must know the sampling distribution of the test statistic when H_0 is true. There is no universal rule for how small a *P*-value in a significance test provides convincing evidence against the null hypothesis.

- ✓ If the *P*-value is smaller than a specified value α (called the **significance level**), the data are **statistically significant** at level α. In that case, we can reject H_0 . If the *P*-value is greater than or equal to α, we fail to reject H_0 .